Radiography Classification: A comparison between Eleven Convolutional Neural Networks

1st Ananda
School of Mathematics, Computer Science and Engineering
City, University of London, London EC1V 0HB, UK
Ananda.Ananda@city.ac.uk, 0000-0001-9537-167X

2nd Cefa Karabağ
School of Mathematics, Computer Science and Engineering
City, University of London, London EC1V 0HB, UK
0000-0003-4424-0471

3rd Aram Ter-Sarkisov
School of Mathematics, Computer Science and Engineering
City, University of London, London EC1V 0HB, UK

4th Eduardo Alonso
School of Mathematics, Computer Science and Engineering
City, University of London, London EC1V 0HB, UK

5th Constantino Carlos Reyes-Aldasoro
School of Mathematics, Computer Science and Engineering
City, University of London, London EC1V 0HB, UK
Constantino-Carlos.Reyes-Aldasoro@city.ac.uk,
0000-0002-9466-2018

***Abstract*—This paper investigates the classification of normal and abnormal radiographic images. Eleven convolutional neural network architectures (*GoogleNet, Vgg-19, AlexNet, SqueezeNet, ResNet-18, Inception-v3, ResNet-50, Vgg-16, ResNet-101, DenseNet-201 and Inception-ResNet-v2*) were used to classify a series of x-ray images from Stanford Musculoskeletal Radiographs (MURA) dataset corresponding to the wrist images of the data base. For each architecture, the results were compared against the known labels (normal / abnormal) and then the following metrics were calculated: accuracy (labels correctly classified) and Cohen's kappa (a measure of agreement) following MURA guidelines. Numerous experiments were conducted by changing classifiers (Adam, Sgdm, RmsProp), the number of epochs, with/without data augmentation. The best results were provided by Inception-Resnet-v2 (Mean accuracy = 0.723, Mean Kappa = 0.506). Interestingly, these results lower than those reported in the Leaderboard of MURA. We speculate that to improve the results from basic CNN architectures several options could be tested, for instance: pre-processing, post-processing or domain knowledge, and ensembles.**

***Index Terms*—CNN, X-ray, Wrist, Classification**

I. Introduction

Wrist fracture is a common injury, especially among older patients [29]. Incidents such as falling, slipping, tripping may lead to fractures that sometimes are ignored by patients and left untreated [33]. The fractures can provoke impairment in the movement of the wrist [3], and in some cases it can lead to serious complications such ruptured tendons or stiffness of the fingers [6]. The basic treatment of fractures, that is, immobilisation and casting, has not changed much in time, as there are Egyptian records describing the re-positioning of bones, fixing with wood and covering with linen [8]. The process of immobilisation is nowadays performed under anaesthesia and thus it is know as Manipulation under Anaesthesia (MUA) and regularly performed in Accidents and Emergency (A&E) departments [4]. The alternative treatment for the fractures is open surgery, which is also known as Open Reduction and Internal Fixation (ORIF) [1]. The surgical procedure is far more complicated than manipulation, and can lead to serious complications [2], however, it is more reliable as a long term treatment as manipulations sometimes fail and then surgery is needed. Despite the considerable amount of research in these areas [1], [2], [4], [5], [13], [22], [23], there is no certainty into which procedure to follow for wrist fractures [14]–[16].

The conditions of the hand and wrist depend on the integrity and function of the ligaments, tendons, muscles, joints, and bones [18]. Imperfect treatment could affect the whole body, causing disruptions at home, work and negatively impact the quality of life [44].

Nowadays, X-ray images have been widely used to visually examine the internal condition of patient abnormalities. The radiologist's interpretation of the X-ray image as a case base clinical information available is a critical point on how the patient is treated [18].

In the United Kingdom, the condition of bone fractures has become an intensive focus, as reflected by the increased demand for diagnostic imaging and intervention radiology [7]. There have been solutions based on not just clinical perspective but also combination with technology works [10], [11], [21], [30], [39].

This work investigates the classification radiographs from the wrist and forearm into two classes, namely normal and abnormal. Traditional analysis of wrists has focused on geometric measurements that are extracted either manually [26], [36], [37], [46] or through image processing [35]. However, in recent times, Artificial intelligence (AI) inspired technology has been used to to tackle some of difficult problems in many areas, among them those related to healthcare and medical

imaging [27], [45]. Thus, in this work we will investigate the performance of eleven different Convolutional Neural Network (CNN) models to assess the classification of wrist fractures into two classes: normal or abnormal.

A large dataset of musculoskeletal radiographs from [34] was used to train to eleven widely-known CNNs. The results, i.e. the ability to distinguish normal and abnormal, provided by different neural networks is the first step to assist a radiologist. Further steps could be directed into deciding the most appropriate treatment for a patient, for instance treat a fracture with MUA or a traditional cast or opt for more complicated and expensive surgery with metal implants.

II. Materials and Methods

A. Materials

This study analysed the wrist radiographs from the public dataset Musculoskeletal Radiographs (MURA) [34]. The dataset has been manually labelled by board-certified radiologists between 2001 and 2012. The radiographs ($n = 14,656$) are divided into images for training ($n = 13,457$), and validation ($n = 1,199$). Furthermore, the radiographs belong to a group called abnormal (i.e. fracture, foreign body, etc.) ($n = 5,818$) or normal ($n = 9,045$). The distribution per anatomical region is shown in Table I and selected cases are illustrated in Fig. 1. Of these, the subset of the **wrists** were selected for this study. In experiments, the actual numbers of data have been checked as it shown in Table II. Furthermore, this study emphasise in classification wrist x-ray images to abnormal and normal category. Table III shows the actual distribution of labelled images in the dataset. Each condition is a combination of data labelled as Valid images and data labelled as Train images.

TABLE I
DISTRIBUTION OF CASES OF THE STANFORD MURA (MUSCULOSKELETAL RADIOGRAPHS) DATA SET [34] FOR STUDIES OF THE UPPER BODY.

No.	Study	Train		Validation		Total
		Normal	Abnormal	Normal	Abnormal	
1	Elbow	1094	660	92	66	1912
2	Finger	1280	655	92	83	2110
3	Hand	1497	521	101	66	2185
4	Humerus	321	271	68	67	727
5	Forearm	590	287	69	64	1010
6	Shoulder	1364	1457	99	95	3015
7	**Wrist**	**2134**	**1326**	**140**	**97**	**3697**
	Total	8280	5177	661	538	14656

B. Convolutional Neural Network architectures

CNNs are a subclass in the hierarchic terminology that includes AI, machine learning, and deep learning [12].

A typical CNN combines a series of layers: convolutional layers followed by sub-sampling layers (Pooling layer), then another convolutional layers followed by pooling layers, and can continue for a certain number of times after which fully-connected layers are added to produce a prediction (e.g. estimated class probabilities). This layer-wise arrangement

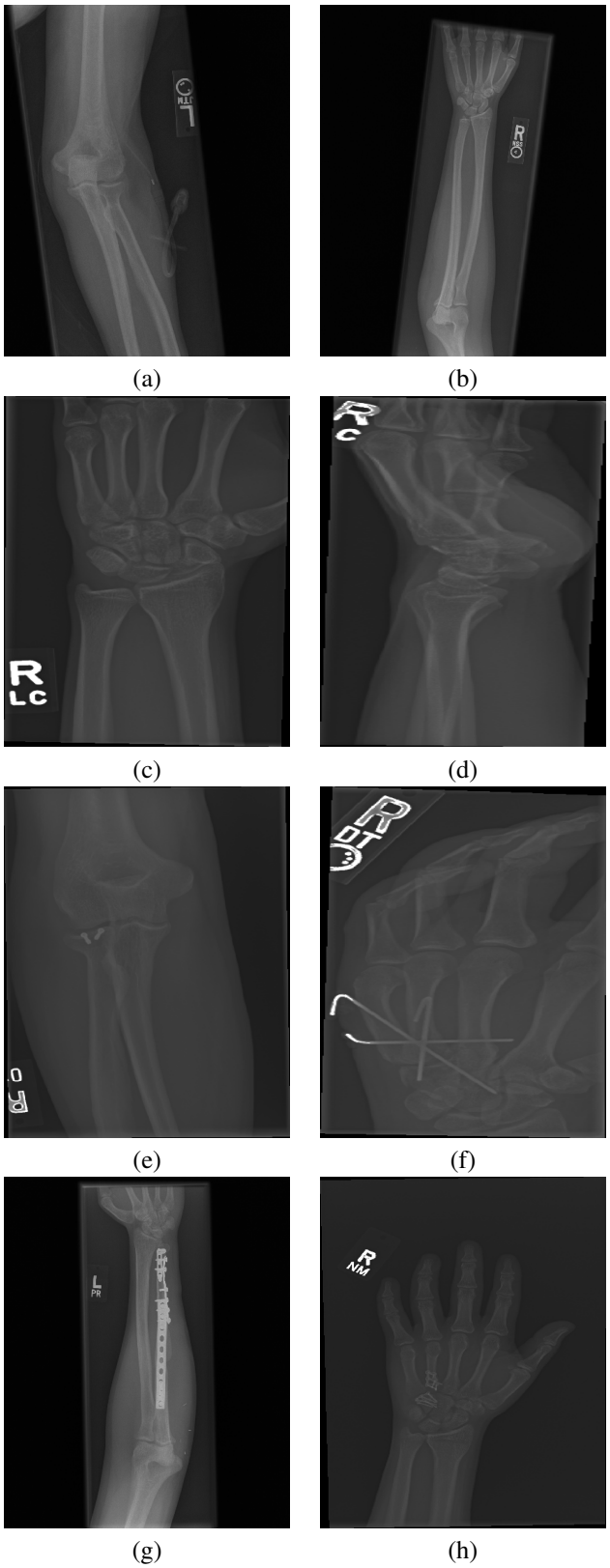

Fig. 1. Illustration of different radiographs of the **MU**sculoskeletal **RA**diographs (MURA) dataset [34] corresponding to training set and negative (no abnormalities) in the top row, and positive (abnormalities) in the bottom row. (a) Elbow, (b) Forearm, (c) Postero-anterior view of Wrist, (d) Lateral view of Wrist, (e) Elbow, (f) Fingers, (g) Forearm, (h) Hand.

TABLE II
DISTRIBUTION OF IMAGES IN MURA (MUSCULOSKELETAL RADIOGRAPHS) DATASET FOR TRAINING AND VALIDATION.

No.	BodyParts	Number Images in Train Folder	Number Images in Valid Folder
1	Elbow	4931	465
2	Finger	5106	461
3	ForeArm	1825	301
4	Hand	5543	460
5	Humerus	1272	288
6	Shoulder	8379	563
7	**Wrist**	**9752**	**659**
Total per condition		36808	3197
Total actual images in dataset		40005	

TABLE III
DISTRIBUTION OF IMAGES IN THE STANFORD MURA (MUSCULOSKELETAL RADIOGRAPHS) DATASET INTO ABNORMAL AND NORMAL GROUPS. THIS WORK CONCENTRATED ON THE WRIST RADIOGRAPHS.

No.	BodyParts	Abnormal (Train + Valid)	Normal (Train + Valid)
1	Elbow	2236	3160
2	Finger	2215	3352
3	ForeArm	812	1314
4	Hand	1673	4330
5	Humerus	739	821
6	Shoulder	4446	4496
7	**Wrist**	**4282**	**6129**
Total per Condition		16403	23602
Images total in dataset		40005	

TABLE IV
WRIST RADIOGRAPHS WERE FURTHER SUBDIVIDED INTO FOUR STUDIES. STUDIES 1,2,3 AND 4 REFER TO A PATIENT VISIT IDENTIFIER, EACH PATIENT MAY HAVE VISITED THE HOSPITAL SEVERAL TIMES. ABNORMAL CORRESPONDS TO POSITIVE ABNORMAL CONDITION LABELLED BY THE EXPERT AND NORMAL CORRESPONDS TO NEGATIVE ABNORMAL CONDITION LABELLED BY THE EXPERT.

Wrist-Train dataset	**Abnormal**	**Normal**
Study 1	3920	5282
Study 2	64	425
Study 3	3	45
Study 4	0	13
Total	3987	5765
Total Wrist Train Images	9752	
Wrist-Valid dataset	**Abnormal**	**Normal**
Study 1	287	293
Study 2	5	59
Study 3	3	9
Study 4	0	3
Total	295	364
Total Wrist Valid Images	659	
Total Images of Wrist	10411	

allows CNNs to combine low-level features to form higher-level features, learn features and eliminate the need for hand crafted feature extractors. In addition, the learned features are translation invariant, incorporate the two-dimensional (2D) spatial structure of images which contributed to CNNs achieving state-of-the-art results in image-related tasks [9].

The input to a CNN, i.e. an image to be classified, transits through the different layers to produce at the end some scores (one score per neuron in the last layer). In the case of image classification, these scores can be interpreted as the probability of the image to belong to each of the classes. The goal of the training process is to learn the weights of the filters at the various layers of the CNN. The output of one of the layers before the last layer, which is fully connected, can be used as a global descriptor for the input image. The descriptor can then be used for various image analysis tasks including classification, recognition, and retrieval [25].

TABLE V
SUMMARY OF CONVOLUTIONAL NEURAL NETWORKS (CNNS) THAT WERE USED IN THIS WORK.

No.	Network	Depth	Image Input Size	Reference
1	GoogleNet	22	224-by-224	[40]
2	Vgg-19	19	224-by-224	[38]
3	AlexNet	8	227-by-227	[24]
4	Squeezenet	18	227-by-227	[20]
5	ResNet-18	18	224-by-224	[17]
6	Inception-v3	48	299-by-299	[42]
7	ResNet-50	50	224-by-224	[17]
8	Vgg-16	16	224-by-224	[38]
9	ResNet-101	101	224-by-224	[17]
10	DenseNet-201	201	224-by-224	[19]
11	Inception-ResNet-v2	164	299-by-299	[41]

The classification of wrist radiographs into two categories (Normal / Abnormal) was considered with eleven CNN architectures. There architectures considered were: **GoogleNet, Vgg-19, AlexNet, SqueezeNet, ResNet-18, Inception-v3, ResNet-50, Vgg-16, ResNet-101, DenseNet-201 and Inception-ResNet-v2**. In addition, the training process of the architecture was tested with different number of epochs (10, 20, 30), different mini-batch sizes (16, 32, 64) and with and without data augmentation. The details of the architectures are displayed in Table V. The experiment pipeline is illustrated in Fig. 2. No pre- or post-processing was applied in any case.

Experiments were conducted in MATLAB®R2018b IDE completed with Deep Learning Toolbox, Image Processing Toolbox and Parallel Computing Toolbox. These experiments were conducted by using a workstation with a processor from INTEL®Xeon ® W-2123 CPU 3.60 GHz, 16GB of 2666MHz DDR4 RAM, 500GB SATA 2.5-inch solid-state drive, and NVIDIA Quadro P620 3GB graphic card.

C. Performance metrics

Accuracy (Ac) was calculated as the proportion of correct predictions among the total number of cases examined, that is:

$$Ac = (TP + TN)/(TP + TN + FP + FN), \qquad (1)$$

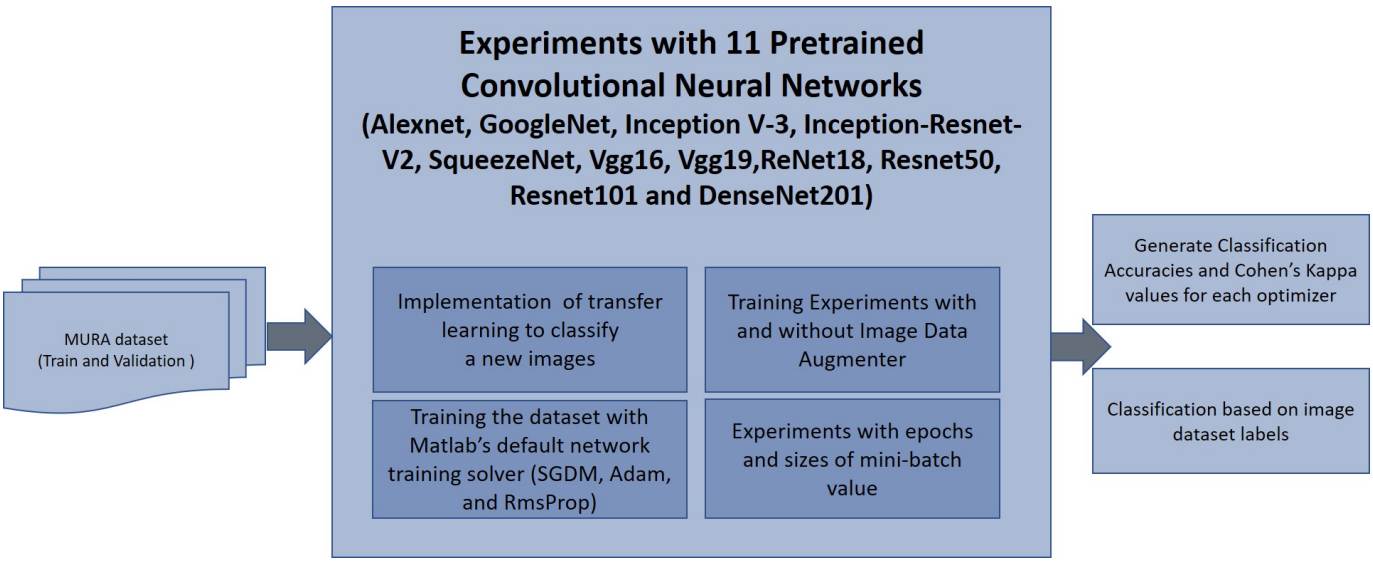

Fig. 2. Block diagram which illustrates the classification of the wrist radiographs with 11 different Convolutional Neural Network (CNN) architectures. 9752 images from **MU**sculoskeletal **RA**diographs (MURA) Wrist dataset were used for training CNN architectures and 659 images were used for validation. Two different metrics, Accuracy (Ac) and Cohen's kappa (κ) were computed to assess the performance of 11 pre-trained CNNs. Image data augmentation was used during training and different number of epochs and mini bach sizes were tested.

where TP and TN correspond to positive and negative classes correctly predicted and FP and FN correspond to false predictions. Cohen's kappa (κ) was also calculated as it is the metric used to rank the MURA challenge [28], [34] and it is considered more robust as it takes into account the possibilities of random agreements. Cohen's Kappa κ was calculated in the following way. With

$$Tot = (TP + TN + FP + FN), \tag{2}$$

being the total number of events, the probability of a yes or TP is

$$P_Y = (TP + FP)(TP + FN)/Tot, \tag{3}$$

the probability of a no, or TN is

$$P_N = (FN + TN)(FP + TN)/Tot, \tag{4}$$

and the probability of random agreement $P_R = P_Y + P_N$, then

$$\kappa = (Ac - P_R)/(1 - P_R). \tag{5}$$

III. Results

Eleven architectures were used to classify the wrist radiograph and the results for accuracy is shown in Table VII and Cohen's Kappa is shown in Table VIII with details of the different classifiers. Each case was tested with different number of epochs (10, 20, 30), different mini-batch sizes (16, 32, 64) and with and without data augmentation. The results in this table aggregate the best results for each architecture.

The results indicated that Inception-Resnet-v2 provided best results both for accuracy ($Ac = 0.723$) and Cohen's kappa ($\kappa = 0.506$). Very close was DenseNet-201 with ($Ac = 0.717$, $\kappa = 0.497$). The lowest results were provided by GoogleNet with ($Ac = 0.654$, $\kappa = 0.381$). Fig. 3 and Fig. 4 illustrate some cases of the classification for Lateral and Postero-anterior views of wrist radiographs.

Not one of the three classifiers provided consistent superiority over the others, and equally, the number of epochs varied and it was not observed that more epochs would consistently provide better results. We speculate that the architectures were converging before these epochs and thus there was not significant advantage of running more epochs. Similarly, not a single option of the size of the mini-batch provided the best results.

IV. Discussion

In this paper, the classification of wrist radiographs with eleven CNN architectures was studied. Whilst Inception-Resnet-v2 provided the best results ($Ac = 0.723$, $\kappa = 0.506$), these were quite far below as compared with the published Leaderboard of MURA, where, at the time of writing (July 2020) the top three results reported $\kappa = 0.843, 0.834, 0.833$ and the best performance for a radiologist was $\kappa = 0.778$. The lowest value of the table is number 70 with $\kappa = 0.518$. Interestingly, many of the architectures here explored appear in the Leaderboard (Inception-ResNet-v2, Vgg-19, DenseNet) and provided higher κ. Whilst in this paper only a subset (the wrists) was explored, it is not considered that the wrists would be more difficult to classify than the other anatomical regions, thus the question arises, *why would all the architectures provide lower results than those at the bottom of the table?* Notice that many variations in the training of the networks, such as epochs, classifiers, size of mini-batch were performed.

We speculate that to improve the results from the standard CNN architectures, such as those analysed in this work, the classification pipelines must include extra steps. Namely:

TABLE VI
SUMMARY OF CONVOLUTIONAL NEURAL NETWORKS (CNNS) HYPERPARAMETERS FOR THIS WORK.

1	GoogleNet	Optimizer	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	64	64	64
		Init. Learn. R.	0.01	0.001	0.001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
2	Vgg-19	Optimizer	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	64	64	64
		Init. Learn. R.	0.001	0.001	0.001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
3	AlexNet	Optimizer	SGDM	ADAM	RMSprop
		Epoch	50	50	50
		Mini batch size	128	128	128
		Init. Learn. R.	0.001	0.001	0.001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
4	SqueezeNet	Optimizer	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	64	64	64
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
5	ResNet-18	Optimizer	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	64	64	64
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
6	Inception-v3	Optimizer	SGDM	ADAM	RMSprop
		Epoch	10	10	10
		Mini batch size	64	64	64
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
7	ResNet-50	Optimizer	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	64	64	64
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
8	Vgg-16	Optimizer	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	128	128	128
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
9	ResNet-101	Optimizer	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	32	32	32
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
10	DenseNet-201	Optimizer	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	32	32	32
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
11	Inception-ResNet-v2	Optimizer	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	32	32	32
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001

TABLE VII
RESULTS OF ACCURACY FOR ELEVEN MODEL OF CONVOLUTIONAL NEURAL NETWORKS USED TO CLASSIFY THE WRIST IMAGES IN MURA DATASET EXPERIMENTS. THE BEST RESULTS FOR EACH ROW ARE HIGHLIGHTED IN *italics* AND THE OVERALL BEST RESULTS ARE HIGHLIGHTED IN **bold**.

No.	CNNs	SGDM	ADAM	Rms Prop	Mean	Ep.	Mini-batch Size
1	GoogleNet	0.650	*0.671*	0.640	0.654	30	64
2	Vgg-19	0.680	*0.681*	0.590	0.650	30	64
3	AlexNet	0.674	*0.690*	0.657	0.674	50	128
4	SqueezeNet	0.683	0.657	*0.690*	0.677	30	64
5	ResNet-18	0.704	*0.709*	0.668	0.693	30	64
6	Inception-v3	*0.710*	0.689	0.707	0.702	10	64
7	ResNet-50	0.686	*0.718*	0.716	0.707	30	64
8	Vgg-16	0.692	0.713	*0.716*	0.707	30	128
9	ResNet-101	*0.715*	0.706	0.701	0.707	30	32
10	DenseNet-201	*0.733*	0.695	0.722	0.717	30	32
11	Inception-ResNet-v2	0.712	***0.747***	0.710	**0.723**	30	32

TABLE VIII
COHEN'S KAPPA RESULTS FROM ELEVEN MODEL OF CONVOLUTIONAL NEURAL NETWORKS USED TO CLASSIFY THE WRIST IMAGES IN MURA DATASET EXPERIMENTS. THE BEST RESULTS FOR EACH ROW ARE HIGHLIGHTED IN *italics* AND THE OVERALL BEST RESULTS ARE HIGHLIGHTED IN **bold**.

No.	CNNs	SGDM	Adam	Rms Prop	Mean	Ep.	Mini-batch Size
1	GoogleNet	0.373	*0.412*	0.358	0.381	30	64
2	Vgg-19	0.433	*0.446*	0.335	0.404	30	64
3	AlexNet	0.420	*0.450*	0.390	0.420	50	128
4	SqueezeNet	0.438	0.390	*0.448*	0.425	30	64
5	ResNet-18	0.474	*0.484*	0.408	0.455	30	64
6	Inception-v3	*0.487*	0.450	0.482	0.473	10	64
7	ResNet-50	0.441	*0.496*	0.494	0.477	30	64
8	Vgg-16	0.453	0.491	*0.492*	0.479	30	128
9	ResNet-101	*0.495*	0.475	0.472	0.481	30	32
10	DenseNet-201	*0.524*	0.458	0.507	0.497	30	32
11	Inception-ResNet-v2	0.485	***0.548***	0.484	**0.506**	30	32

1) **Pre-processing steps**, which may consist of: Low pass filtering to remove high-frequency noise, cropping of images to remove excessive background region (notice that some of the incorrect classifications in Fig. 3 had large background regions). More elaborate pre-processing approaches such as location and orientation of bones [35] could help detect the areas of real interest, and discard any region that may be biasing results, such as the labels for right or left hand, which being always very bright might be confusing the architectures.
2) **Post-processing steps** may also be considered, for instance, the association between key features and the predicted classes [32], [43]. Furthermore, the visualisation of key features may be useful to stakeholders (e.g. clinicians or radiologists) who might be more interested in the attributes of the original data rather than the

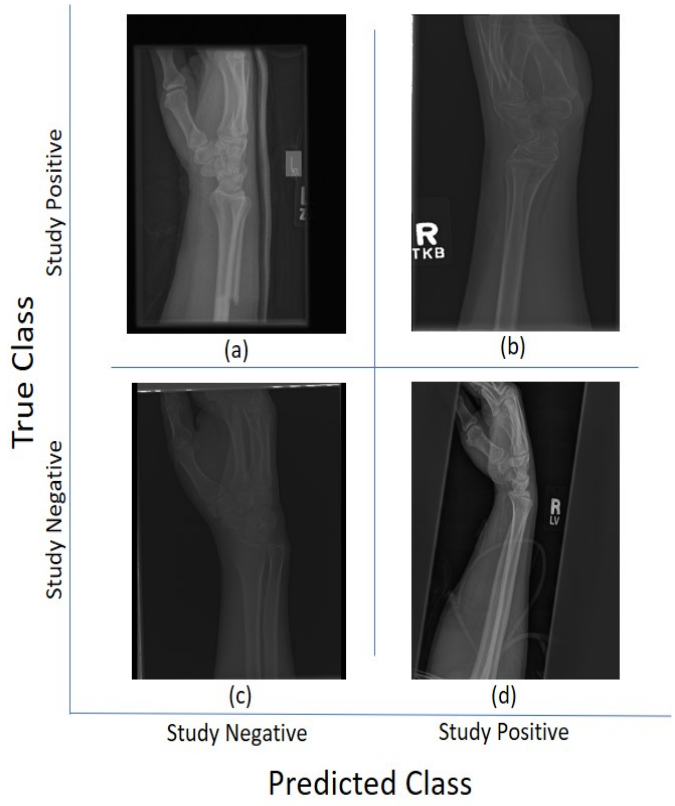

Fig. 3. Illustration of classification results for Lateral (LA) views of wrist radiographs. (a) Corresponds to positive (abnormal) diagnosis image but predicted as negative (normal), (b) Abnormal diagnosis and abnormal prediction. (c) Normal diagnosis image and normal prediction. (d) Normal diagnosis and abnormal prediction. Notice that the errors in classification may have been biased by artefactual elements on the images.

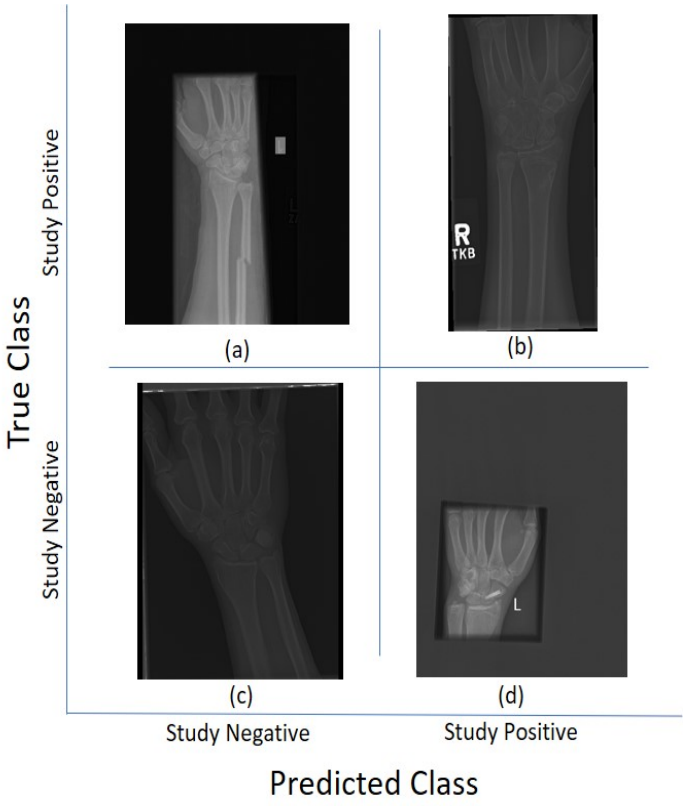

Fig. 4. Illustration of classification results for Postero-Anterior (PA) views of wrist radiographs. (a) Corresponds to positive (abnormal) diagnosis image but predicted as negative (normal), (b) Abnormal diagnosis and abnormal prediction. (c) Normal diagnosis image and normal prediction. (d) Normal diagnosis and abnormal prediction. Notice again that the errors in classification may have been biased by artefactual elements on the images.

architectures themselves [31].

3) **Ensembles** or combination of different configurations may also help increase the results of individual configurations.
4) Finally, adding **domain knowledge** in terms of knowledge of the anatomical region (i.e. elbow or hand) with the possible cases (i.e. fracture or implant) may allow the fine tuning of the architectures to detect not only an abnormality but the type of abnormality and the location of this.

V. Conclusion

In this paper, we described a comparison of eleven convolutional neural networks to classify normal and abnormal radiographic images. The results obtained did not represent a striking value and required additional steps in the classification process. We suggest further steps such as image noise removal, reduction of excessive background region, image features association with predicted classes, enhanced visualisation of a classified image, variations of training configuration, and addition of anatomical region into the data set.

The classification of an image to normal or abnormal conditions is not only restricted to medical domains but also could apply to other areas such as inspection of production defects from manufacturing products, animal treatments in the veterinary medicines domain, and object detection from satellite imagery such as landforms, settlements, and other geographical features.

References

[1] R. Arora, M. Gabl, M. Gschwentner, C. Deml, D. Krappinger, and M. Lutz, "A Comparative Study of Clinical and Radiologic Outcomes of Unstable Colles Type Distal Radius Fractures in Patients Older Than 70 Years: Nonoperative Treatment Versus Volar Locking Plating," *Journal of Orthopaedic Trauma*, vol. 23, no. 4, pp. 237–242, Apr. 2009.

[2] R. Arora, M. Lutz, A. Hennerbichler, D. Krappinger, D. E. Md, and M. Gabl, "Complications Following Internal Fixation of Unstable Distal Radius Fracture With a Palmar Locking-Plate," *Journal of Orthopaedic Trauma*, vol. 21, no. 5, pp. 316–322, May 2007.

[3] R. W. Bacorn and J. F. Kurtzke, "COLLES' FRACTURE: A Study of Two Thousand Cases from the New York State Workmen's Compensation Board," *JBJS*, vol. 35, no. 3, pp. 643–658, Jul. 1953.

[4] A. Barai, B. Lambie, C. Cosgrave, and J. Baxter, "Management of distal radius fractures in the emergency department: A long-term functional outcome measure study with the disabilities of arm, shoulder and hand (dash) scores," *Emergency Medicine Australasia*, vol. 30, no. 4, pp. 530–537, 2018.

[5] C. Bartl, D. Stengel, T. Bruckner, I. Rossion, S. Luntz, C. Seiler, and F. Gebhard, "Open reduction and internal fixation versus casting for highly comminuted and intra-articular fractures of the distal radius

(ORCHID): protocol for a randomized clinical multi-center trial," *Trials*, vol. 12, no. 1, p. 84, Mar. 2011.

[6] W. P. Cooney, J. H. Dobyns, and R. L. Linscheid, "Complications of Colles' fractures," *The Journal of Bone and Joint Surgery. American Volume*, vol. 62, no. 4, pp. 613–619, 1980.

[7] E. M. Curtis, R. van der Velde, R. J. Moon, J. P. W. van den Bergh, P. Geusens, F. de Vries, T. P. van Staa, C. Cooper, and N. C. Harvey, "Epidemiology of fractures in the united kingdom 1988-2012: Variation with age, sex, geography, ethnicity and socioeconomic status," *Bone*, vol. 87, pp. 19–26, Jun 2016.

[8] R. J. Diaz-Garcia and K. C. Chung, "The evolution of distal radius fracture management – a historical treatise," *Hand Clinics*, vol. 28, no. 2, pp. 105–111, May 2012.

[9] R. Ebsim, J. Naqvi, and T. F. Cootes, "Automatic detection of wrist fractures from posteroanterior and lateral radiographs: A deep learning-based approach," in *International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging*. Springer, 2018, pp. 114–125.

[10] M. P. Gaspar, J. Lou, P. M. Kane, S. M. Jacoby, A. L. Osterman, and R. W. Culp, "Complications following partial and total wrist arthroplasty: A single-center retrospective review," *Journal of Hand Surgery*, vol. 41, no. 1, pp. 47–53.e4, Jan 2016.

[11] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, and et al., "Niftynet: a deep-learning platform for medical imaging," *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 113–122, May 2018.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[13] R. Grewal, J. C. MacDermid, G. J. W. King, and K. J. Faber, "Open Reduction Internal Fixation Versus Percutaneous Pinning With External Fixation of Distal Radius Fractures: A Prospective, Randomized Clinical Trial," *Journal of Hand Surgery*, vol. 36, no. 12, pp. 1899–1906, Dec. 2011.

[14] H. H. Handoll, J. S. Huntley, and R. Madhok, "Different methods of external fixation for treating distal radial fractures in adults," *Cochrane Database of Systematic Reviews*, no. 1, 2008.

[15] H. H. Handoll and R. Madhok, "Closed reduction methods for treating distal radial fractures in adults," *Cochrane Database of Systematic Reviews*, no. 1, 2003.

[16] ——, "Conservative interventions for treating distal radial fractures in adults," *Cochrane Database of Systematic Reviews*, no. 2, 2003.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016, pp. 770–778.

[18] H. Hsu, M. P. Fahrenkopf, and S. V. Nallamothu, *Wrist Fracture*. StatPearls Publishing, 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK499972/

[19] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *arXiv:1608.06993 [cs]*, Aug 2016, arXiv: 1608.06993. [Online]. Available: http://arxiv.org/abs/1608.06993

[20] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size," *arXiv:1602.07360 [cs]*, Feb 2016, arXiv: 1602.07360. [Online]. Available: http://arxiv.org/abs/1602.07360

[21] J. Iglesias and M. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Medical Image Analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[22] H. Kapoor, A. Agarwal, and B. K. Dhaon, "Displaced intra-articular fractures of distal radius: a comparative evaluation of results following closed reduction, external fixation and open reduction with internal fixation," *Injury*, vol. 31, no. 2, pp. 75–79, Mar. 2000.

[23] A. J. Kelly, D. Warwick, T. P. K. Crichlow, and G. C. Bannister, "Is manipulation of moderately displaced Colles' fracture worthwhile? A prospective randomized trial," *Injury*, vol. 28, no. 4, pp. 283–287, May 1997.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*. Curran Associates, Inc., 2012, p. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[25] H. Laga, Y. Guo, H. Tabia, R. Fisher, and M. Bennamoun, *3D Shape Analysis: Fundamentals, Theory, and Applications*. United States: Wiley-Blackwell, 2019.

[26] J. I. Lee, K. C. Park, I.-H. Joo, H. W. Jeong, and J. W. Park, "The effect of osteoporosis on the outcomes after volar locking plate fixation in female patients older than 50 years with unstable distal radius fractures," *The Journal of Hand Surgery*, vol. 43, no. 8, p. 731–737, 2018.

[27] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big data application in biomedical research and health care: A literature review," *Biomedical Informatics Insights*, vol. 8, p. BII.S31559, Jan 2016.

[28] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, vol. 22, no. 3, p. 276–282, Oct 2012.

[29] S. Meena, P. Sharma, A. K. Sambharia, and A. Dawar, "Fractures of distal radius: An overview," *Journal of Family Medicine and Primary Care*, vol. 3, no. 4, pp. 325–332, 2014.

[30] P. Meyer, V. Noblet, C. Mazzara, and A. Lallement, "Survey on deep learning for radiotherapy," *Computers in Biology and Medicine*, vol. 98, pp. 126–146, Jul 2018.

[31] K. H. Ngan, A. d. Garcez, K. M. Knapp, A. Appelboam, and C. C. Reyes-Aldasoro, "Making densenet interpretable, a case study in clinical radiology," *medRxiv*, p. 19013730, Dec 2019.

[32] J. Oramas, K. Wang, and T. Tuytelaars, "Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=H1ziPjC5Fm

[33] N. Raby, L. Berman, S. Morley, and G. De Lacey, *Accident and Emergency Radiology: A Survival Guide (Third Edition)*. Saunders Elsevier, 2015.

[34] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, and et al., "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," Dec 2017. [Online]. Available: http://arxiv.org/abs/1712.06957

[35] C. C. Reyes-Aldasoro, K. H. Ngan, A. Ananda, A. d. Garcez, A. Appelboam, and K. M. Knapp, "Geometric semi-automatic analysis of colles' fractures," *medRxiv*, p. 2020.02.18.20024562, Feb 2020.

[36] S. H. Rhee and J. Kim, "Distal radius fracture metaphyseal comminution: A new radiographic parameter for quantifying, the metaphyseal collapse ratio (mcr)," *Orthopaedics & Traumatology: Surgery & Research*, vol. 99, no. 6, p. 713–718, Oct 2013.

[37] B. Sharareh and S. Mitchell, "Radiographic outcomes of dorsal spanning plate for treatment of comminuted distal radius fractures in non-elderly patients," *Journal of Hand Surgery Global Online*, vol. 2, no. 2, p. 94–101, Mar 2020.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556 [cs]*, Apr 2015, arXiv: 1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[39] S. Siuly and Y. Zhang, "Medical big data: Neurological diseases diagnosis through medical data analysis," *Data Science and Engineering*, vol. 1, no. 2, pp. 54–64, Jun 2016.

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015, p. 1–9.

[41] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv:1602.07261 [cs]*, Aug 2016, arXiv: 1602.07261. [Online]. Available: http://arxiv.org/abs/1602.07261

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *arXiv:1512.00567 [cs]*, Dec 2015, arXiv: 1512.00567. [Online]. Available: http://arxiv.org/abs/1512.00567

[43] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," *ArXiv*, 2018. [Online]. Available: http://arxiv.org/abs/1808.01974

[44] I. Vergara, K. Vrotsou, M. Orive, S. Garcia-Gutierrez, N. Gonzalez, C. Las Hayas, and J. M. Quintana, "Wrist fractures and their impact in daily living functionality on elderly people: a prospective cohort study," *BMC geriatrics*, vol. 16, p. 11, Jan 2016.

[45] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: Big data for personalized healthcare," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1209–1215, Jul 2015.

[46] J. Wang, Y. Lu, Y. Cui, X. Wei, and J. Sun, "Is volar locking plate superior to external fixation for distal radius fractures? a comprehensive meta-analysis," *Acta Orthopaedica Et Traumatologica Turcica*, vol. 52, no. 5, p. 334–342, Sep 2018.